Automatic Generation of Quantitative Reasoning Items

A Pilot Study

Martin Arendasy¹, Markus Sommer², Georg Gittler¹, and Andreas Hergovich³

¹Differential and Personality Research Group, Faculty of Psychology, University of Vienna, ²Dr. Schuhfried GmbH, Mödling, ³Social Psychology Research Group, Faculty of Psychology, University of Vienna, all Austria

Abstract. This paper deals with three studies on the computer-based, automatic generation of algebra word problems. The cognitive psychology based generative/quality control frameworks of the item generator are presented. In Study I the quality control framework is empirically tested using a first set of automatically generated items. Study II replicates the findings of Study I using a larger set of automatically generated algebra word problems. Study III deals with the generative framework of the item generator by testing construct validity aspects of the item generator produced items. Using nine Rasch-homogeneous subscales of the new intelligence structure battery (INSBAT, Hornke et al., 2004), a hierarchical confirmatory factor analysis is reported, which provides first evidence of convergent as well as divergent validity of the automatically generated items. The end of the paper discusses possible advantages of automatic item generation in general ranging from test security issues and the possibility of a more precise psychological assessment to mass testing and economical questions of test construction.

Keywords: algebra word problems, automatic item generation, quantitative reasoning, Rasch model

Introduction

It is becoming increasingly common for psychological tests to be used in the selection of suitable job applicants (Schu- ler, 2000) or in career counseling. However, this also comes with an increased exposure of test content to the public. Thus, aspects of test security have become a more and more important issue in the past several decades, leading to methodological developments that aim to increase test security. Examples of such methods are (1) the use of computerized adaptive tests (CAT; Van der Linden & Glas, 2000; Wainer, 2000), (2) the implementation of item exposure control algorithms (e.g., Harris & Jackson, 2002; Stocking & Lewis, 2000; Sympson & Hetter, 1985), (3) the use of special software packages that prevent cheating in online-assessments (Edelblut, Elliot, Mikulas, & Bosley, 2002), and (4) the development of various methods of automatic item generation (AIG; for an introductory overview see Irvine & Kyllonen, 2002).

According to Bejar (2002), the various approaches to AIG can be distinguished with regard to their level of generativity, which subsumes (1) the degree of theoretical foundation of the item generator, and (2) the degree of automation of the item generation process. Functional item generation is characterized by the lack of a cognitive model, while model-based item generation is based on a cognis-
proach has been used by the Educational Testing Service (ETS) to generate quantitative reasoning items for the graduate research examination (GRE) in order to extend existing item pools.

Even though automatic item generation is a comparatively new field in Applied Psychometrics, its theoretical foundations are quite advanced. However, a question that has received little attention thus far is whether the generative framework has to be extended by an (automated) quality control framework in order to be able to generate items at a high psychometric level (Arendasy, 2004, 2005, in press). The generative framework basically includes a set of radicals derived from cognitive psychology based models of the latent trait to be measured (cf. Arendasy & Sommer, 2005; Arendasy, Sommer, & Ponocny, 2005) and serves as a specification of the construct-related solution strategies. In line with Greeno, Smith, and Moore (1993), and Siegler (1996), Arendasy (2005, in press) assumes that a test person’s solution strategy choice is not merely a characteristic of the test person himself, but arises from an interaction between characteristics of the respondent and characteristics of the item material. Different design features of the item material may suggest different solution strategies that can be more or less construct-related. Hence, the item material should be designed in a manner that provides an “affordance” (Greeno et al., 1993) for solution strategies to be related to the latent trait to be measured. Technically, the affordance of item material can be defined by the set radicals implemented into the generative framework of an item generator. At the same time one needs to take care that alternate solution strategies leading to differential item functioning (Holland & Wainer, 1993) are not supported by the test material. This latter goal can be achieved by defining a set of constraints (Greeno et al., 1993) that subsequently has to be implemented into the quality control framework of an item generator. The constraints can be derived from prior studies on possible reasons of differential item functioning (for examples see Arendasy & Gittler, 2003) or psychometric experiments specifically designed to investigate the impact of constraints implemented in a quality control framework (for recent examples of this approach to the ability domains “figural reasoning” or “spatial ability” see Arendasy & Gittler, 2003; Arendasy & Sommer, 2005). With regard to the psychometric quality notion of Rasch-homogeneity has to be mentioned, which directly leads to the essential methodological background of this paper.

Methodological Background

The item response theory (IRT; cf. Van der Linden & Hambleton, 1997) offers a measurement model that – apart from other advantages – enables one to investigate the dimensionality of test items in a theoretically sound way. The homogeneity problem of dichotomous items can be analyzed within the framework of IRT with the assistance of, e.g., the Rasch model (RM; Rasch, 1960/1980). If the RM holds true for a defined set of items and for particular groups of persons, the test is said to be “Rasch-homogeneous.” Rasch-homogeneous tests have a number of favorable psychometric characteristics (for details see Embretson & Reise, 2000; Fischer & Molenaar, 1995) that set them apart from scale constructions based on classical test theory (CTT). Two of the most important practical advantages are (1) equality of item parameter estimates over different subgroups of respondents (commonly referred to as the “person-homogeneity”-assumption of the RM; see Rost, 2004) and (2) the one-dimensionality of Rasch-calibrated scales (sometimes referred to as the “item-homogeneity”-assumption of the RM; Rost, 2004).

The model fit of the RM can be assessed using different approaches, e.g., by well known, classical likelihood-ratio tests, which relate the likelihood of various restrictive model estimates to each other and transfer these estimates to an asymptotically \( \chi^2 \)-distributed statistic. However, such statistics rely on some sort of parameter estimation (Glas & Verhelst, 1995). This means that they are only accurate given a sufficient sample size, which is not always available in practical research (Glas & Verhelst, 1995; Ponocny, 2001). One way to circumvent this difficulty is to use nonparametric goodness of fit statistics, which do not rely on parameter estimation. Ponocny (2001) recently introduced a family of nonparametric goodness of fit statistics that are conditional on the item and subject marginals of data matrices. In essence these goodness of fit statistics resort to a Monte-Carlo shuffling algorithm that exchanges the dichotomous entries in the tetrades of a data matrix while retaining the item and subject marginals. This idea is based on the assumption that all data matrices with given item and subject marginals are equally likely in case of a fit of the RM. Ponocny (2001) developed several T-statistics, which essentially determine the proportion of certain data matrices on matrices, which are in accordance with the RM (see also Rost, 2004). The fit of a given data matrix to the Rasch model is determined by means of a p-value. When a p-value is above .05 one can assume that the investigated model assumption can be retained. The benefit of these statistics over a parametric approach resides in the fact that these statistics constitute powerful tests of a given model assumption (e.g., person homogeneity, item homogeneity), which do not depend on the sample size but on the number of simulations carried out with the Monte-Carlo shuffling algorithm (cf. Ponocny, 2001, p. 445). Thus, one may speak of exact tests, because the approximation can be made theoretically arbitrarily precise (Ponocny, 2001). In order to test the assumption of global person homogeneity of the RM, the statistic \( T_{10} \) can be used; this nonparametric approach to the investigation of the assumption of person homogeneity enables the formulation of specific hypotheses regarding the items responsible for an assumed misfit, which increases the power of the test statistic (Ponocny, 2001). The assumption of item homogeneity can be
checked using the statistic $T_2$. Both statistics essentially relate the number of simulated matrices that are in accordance with the RM assumption under investigation to the total number of simulated matrices.

An additional and elegant way of testing the RM-fit is provided by means associated with fit indices for the mixed Rasch model (MRM; Rost, 2004). The MRM is less restrictive and can be seen as an extension of the RM. From a mathematical point of view, the MRM constitutes a generic model that integrates latent class analysis and the RM and, thus, includes a qualitative and a quantitative component. The model assumes that one-dimensionality doesn’t hold for the sample as a whole but only for qualitatively different classes within the sample. These qualitative differences are reflected either in the order of or the differences between the item parameters. The model equation of the MRM is similar to the RM equation extended by the class membership $g$. The probability of a subject $s$ belonging to Class $g$ to solve Item $i$ is specified in equation 1.2.

$$P(X_{si} = 1 | \theta_{s}, \beta_{i,g}) = \frac{\exp(\theta_{sg} - \beta_{ig})}{1 + \exp(\theta_{sg} - \beta_{ig})} \quad [1.1]$$

As can be seen in equation 1.1, the probability of a correct answer depends on a class-specific latent-trait parameter (e.g., ability) as well as class-specific item parameters. The MRM can easily be transformed into the RM by assuming that the number of latent classes is 1. This means that a one-class solution is equivalent to the fit of the RM (Rost, 2004). The main advantage of this approach to RM-fit-testing is that, if a one-class solution fits the empirical data, no partitioning criteria exist for which the RM does not hold true. For practical aspects of the application of the MRM see Rost (2004).

**Application to the Measurement of “Quantitative Reasoning”**

The theoretical considerations about the construction of item generators outlined above are applied to the automatic generation of quantitative reasoning items using an isomorph approach. Quantitative reasoning constitutes a second-order factor within the modified $G_1$-$G_2$ theory proposed by Horn (1989; Horn & Noll, 1997) as well as the latest version of the three-stratum theory of intelligence (Bickley, Keith, & Wolfe, 1995; Carroll, 1993). This second-order factor has been consistently identified in several studies investigating the structure of intelligence (e.g., Bickley et al., 1995; Carroll, 1993; Gustafsson, 1984; Horn & Noll, 1997; McGrew & Woodcock, 2001; Roid, 2003; Undheim & Gustafsson, 1987) and proved to be a significant predictor of academic success (Ackerman, 1989). According to Horn and Noll (1997) quantitative reasoning comprises respondents’ number sense as well as their understanding of basic arithmetic or algebraic principles and their ability to apply them in a problem-solving context.

In the past decades several researchers have investigated the cognitive processes used to solve algebra word problems (e.g., Hall, Kibler, Wenger, & Truxaw, 1989; Hinsley, Hayes, & Simon, 1977; Kintsch, 1998; Koedinger & Nathan, 2004; Lane, 1991; Mayer, Larkin, & Kadane, 1984; Nathan, Kintsch, & Young, 1992; Reusser, 1989; Tirre & Pena, 1993) as well as the item features influencing their difficulty (e.g., Enright, Morley, & Sheehan, 2002; Koedinger & Nathan, 2004; Lane, 1991; Lepik, 1990; Reed, 1987; Sebrecht, Enright, Bennett, & Martin, 1996; Shalin & Bee, 1985; Yerushalmi & Gilead, 1999). The results of these studies provide valuable insights for the automatic generation of algebra word problems.

According to Sebrecht et al. (1996) there are two primary sources of item difficulty in algebra word problems: working memory (Baddeley, 1992) and mathematical as well as real-world knowledge. During the solution of algebra word problems the mathematical and real-world knowledge retrieved from long-term memory has to be reconciled with the information from the problem text in working memory to derive a solution plan (Koedinger & Nathan, 2004). Based on the literature on algebra word problem solving, the solution process can be divided into five phases (Kintsch, 1998; Koedinger & Nathan, 2004; Mayer et al., 1984).

In the first phase, respondents resort to their linguistic knowledge to interpret what is being stated in the given problem text. This step is often referred to as problem translation (Mayer et al., 1984). The result is an internal representation of the problem (“text base”; Kintsch, 1998) the construction of which is most likely associated with non-quantitative reasoning-related crystallized intelligence (Horn, 1989; Horn & Noll, 1997). As a consequence, “grammatical and linguistic complexity” has to be restricted by a corresponding constraint in the quality control framework to ensure that uncommon phrases and grammatically complex formulations are not used by the item generator.

In the second phase the respondents start to elaborate the text base by retrieving mathematical and real-world knowledge using problem schemata from long-term memory. According to Mayer (1981) there are different schemata for various kinds of algebra word problems. In order to avoid the individual differences in respondents’ general word problems, schema (e.g., Reusser & Stebler, 1997) exert a differential effect on the psychometric item properties, and crucial information, such as the assumption of constant rates, has to be made explicit in the word problem text. This feature can be regarded as an essential constraint that ensures the unambiguousness of the problem text.

According to Mayer (1981) each algebra word-problem family can be characterized by a unique set of basic schemata (e.g., speed = distance/time). Based on this theoretical consideration we assumed that the use of a more varied set of algebra word-problem families and their various sub-
types could result in a violation of the assumption of item homogeneity. Thus, a constraint regarding the algebra word-problem types was implemented, which restricted the generation process to various types of “rate problems,” excluding in advance other algebra word-problem types, such as geometry and probability word problems.

According to Bernardo (1994), the algebra word problem schemata derived from long-term memory also contain information about their typicality, which can be used to facilitate their retrieval. Bernardo (1994), Blessing and Ross (1996), and Enright et al. (2002) have been able to show that the typicality of the cover story – measured as the relative frequency with which a certain cover story is used for a certain algebra word problem type (Mayer, 1981) – influences the item difficulty of a given algebra word problem. This item feature can then be used as a radical by integrating it into the generative framework of the item generator.

The new information derived from a problem schema needs to be reconciled with the information from the text base. The result of the so-called problem integration phase (Kintsch, 1998) is a situational model (Kintsch, 1998; Nathan et al., 1992; Reusser, 1989) that represents the respondents’ understanding of the situation described in the word problem text. Theoretically the construction of the situation model could be complicated by altering the “presentational structure” (e.g., Koedinger & Nathan, 2004; Staub & Reusser, 1995) of the information presented in the word problem text. However, this alteration would result in a possible source of variance in the item difficulty parameters, which is associated with respondents’ text comprehension ability instead of their quantitative reasoning ability. Thus, a constraint has to be implemented in the quality control framework that fixes the presentational structure in a manner to ensure all word problems are presented in an “order naturalis” (Staub & Reusser, 1995) by representing respondents’ preferred information presentation order (which is in line with the actual time frame of the events described in the problem text).

The third phase refers to the actual mathematization process. The situation model serves as a starting point for the construction of the mathematical problem model (Reusser, 1989) by providing the functional constraints that need to be kept in mind during the mathematization process (Kintsch, 1998; Nathan et al. 1992). According to Hall et al. (1989; Nathan et al., 1992; Koedinger & Nathan, 2004) this phase constitutes the core of “quantitative reasoning” (Horn, 1989; Horn & Noll, 1997). Erroneous mathematical problem models mainly result from ignoring central affordances and constraints in the situation model (Reusser, 1989). The number of functional constraints and affordances in a given algebra word problem can be quantified by a detailed analysis of the episodic (Shalin & Bee, 1985; Hall et al., 1989) or algebraic (Enright et al., 2002; Reed, 1987) structure of the given algebra word problem. Both procedures result in a number of functional relations between the elements of an algebra word problem (Hall et al., 1989; Shalin & Bee, 1985) corresponding to the number of partial equations that can be set up for these relations (Enright et al., 2002; Reed, 1987). Sebrecht et al. (1996), Lane (1991), and Enright et al. (2002) have been able to demonstrate that the number of partial equations significantly affects the difficulty of an algebra word problem. This item feature is integrated as a radical in the generative framework.

In the fourth phase the respondents derive a solution plan from the mathematical problem model (Reusser, 1989) using their strategic knowledge. This solution phase is commonly referred to as solution planning and monitoring (Mayer et al., 1984). Respondents may resort to various solution strategies in order to solve a given algebra word problem. These solution strategies can roughly be divided into formal-algebraic solution strategies, which are characterized by a translation of the mathematical problem model into a solution equation (e.g., Bobrow, 1968; Nhouyvanisong, 1999; Paige & Simon, 1966) and informal solution strategies, which re-sort primary to model-based reasoning based on the respondents’ situational understanding of the algebra word problem (e.g., Hall et al., 1989; Katz, Friedman, Bennett, & Berger, 1996; Koedinger & Nathan, 2004; Nathan et al., 1992; Nhouyvanisong, 1999). In contrast to formal-algebraic solution strategies, informal strategies do not require such a mathematical formalization. These two types of solution strategies are associated with different demands on respondents’ understanding of the principles of algebra. Research conducted by Hall et al. (1989), Katz et al. (1996), Koedinger and Nathan (2004), and Tabachneck et al. (1994) indicates that formal-algebraic solution strategies are seldom used; most respondents predominantly resort to informal solution strategies when solving algebra word problems. Only when faced with more complex algebra word problems are formal-algebraic solution strategies commonly used. The algebraic complexity of a given algebra word problem, thus, constitutes another radical for the automatic generation of algebra word problem. According to Enright et al. (2002), the algebraic complexity can be measured by the number of unknown elements and the number of embedded unknown elements in a solution-relevant equation. These authors also have been able to demonstrate that such a measure of complexity is significantly related to the item difficulty parameter of a three-parameter Birnbaum model. The total number of unknown elements in the solution-relevant equation is integrated as a further radical in the generative framework.

In the final phase the respondents calculate the solution to a given algebra word problem using their knowledge of arithmetic procedures. This solution step is commonly referred to as solution execution (Mayer et al., 1984).

Based on the theoretical considerations regarding automatic item generation and algebra word-problem solving outlined above, the aim of the present study was the construction of an item generator for algebra word problems using a schema-based approach and the empirical evaluation of its products. In line with the theoretical consider-
ations outlined above, the item generator consists of a combination of quality control and generative frameworks to ensure the construction of algebra word problem items at a well-defined psychometric level (see Methods below).

Method

General Description of the Item Generator and Its Products

Based on the theoretical model outlined above a set of constraints and radicals was defined and implemented into the quality control and generative frameworks of the item generator. Unlike item generators using an item-variants approach, the constraints and radicals are embedded into the templates of the rate word-problem types described by Mayer (1981). The technical construction of the items was done using the item generator AGen (Arendasy, 2004). Each template in the item generator consists of a set of mandatory and optional sentences presented in an “order naturalis” (Kintsch, 1998; Staub & Reusser, 1995), whereby the optional sentences are used to increase either the number of partial equations or the number of unknown elements in the solution equation or both. Technically speaking, AGen could be considered a combination of an item-variant approach and an item-isomorph approach (Dennis et al., 2002). The ambiguity of the templates was checked by an expert in German philology and an experienced test instruction designer. When both experts agreed that the item template was unambiguous, the template was included in the item generator. Furthermore, the uniqueness of the solution for each algebra word problem generated was ensured by means of a rule scanner, which checks that any item generated can only be accurately solved by a single equation or set of equations (see Arendasy, in press). The rule scanner, thus, is an additional constraint implemented in the item generator.

Table 1 provides an overview of the constraints and radicals implemented in the item generator. The actual numerical values as well as the variable parts of the cover story, such as precise objects, etc., served as incidentals.

AGen generates algebra word problems using the radicals and constraints outlined in Table 1. Furthermore, the item generator generates an item-specific protocol that includes all relevant item characteristics as well as the parameters derived from them.

General Procedure

In order to investigate the products of the new item generator AGen, a series of studies investigating the dimensionality, reliability, and construct validity were carried out. For the first two studies, the test was administered using the TestWeb (Arendasy, 2002) which enables web-based psychological assessment. The first two studies were carried out during university seminars held by the first author in small groups of no more than three persons. Each graduate student test administrator was personally present during the administration of the test. The respondents were instructed that they would see several algebra word problems and should find out the numerical value of the unknown element. The test used a free-response item format to ensure that guessing would not affect the estimation of the item parameters in the way a multiple choice format could. All algebra word-problem items were presented in a power setting without any time limitations. The respondents were able to correct their answers until they pressed the “next” button to reach the next item of the test. They were not allowed to go back to an already solved item to correct their answer.

The third study was conducted using the Vienna Test System (VTS). In this study the items used in Study I were administered as a subscale of the new intelligence structure battery (INSBAT; Hornke et al., 2004), which is based on the modified Gf-Gc theory proposed by Horn (1989; Horn & Noll, 1997). In addition, nine other subscales as indicators for the stratum-two factors Gq (quantitative reasoning), Gc (crystallized intelligence), Gf (fluid intelligence), Gltm (longterm memory), and Gstm (short-term memory) were administered to investigate the nomothetic span (Embreton, 1983) of the new algebra word-problem test described in this paper.

Study I

Study I was designed to examine the necessity of a constraint regarding the restriction of the algebra word-problem families implemented into the quality control framework of the item generator. According to Mayer (1981) different types of algebra word-problem families require different basic schemata to be solved. We, thus, assumed that the necessity to resort to different schemata could result into a violation of the assumption of item homogeneity. In order to investigate this assumption, a total of k = 8 rate word problems was generated using a set of four radicals outlined in Table 1. Furthermore, k = 8 probability word problems were generated using the same set of radicals with slight adjustments to this specific problem type (see also Enright et al., 2002). These k = 16 algebra word problems were presented together in a predefined, fixed item order that was determined randomly. The instruction given to the respondents in this study was identical to the one outlined in the description of the general procedure.

Sample

The sample consisted of 46 (46.5%) male and 53 (53.5%) female respondents aged 18 to 45 years (Mean=28.51; SD
A total of 11 (11.1%) respondents had completed 9 or 10 years of schooling without completing professional training (ISCED education level 2), 32 (32.3%) respondents had completed professional training or technical school (ISCED 3), 39 (39.4%) respondents had graduated from high-school (ISCED 4), and 17 (17.2%) respondents had graduated from college or university (ISCED 5).

Results

In order to investigate whether rate-type algebra word problems and probability-type algebra word problems violate the assumption of person homogeneity we calculated various $T_{10}$ statistics for the combined scale. The calculations were done with the software program T-Rasch (Ponocny & Ponocny-Seliger, 1999). Person homogeneity was assessed using the partitioning criteria rawscore, sex, age, and educational level. In this analysis the $\alpha$-level was set to .01 in advance because of the number of model tests. The results of the $T_{10}$ statistics are presented in Table 2.

The empirical $p$-value for the partitioning criteria educational-level is below $\alpha = .01$, while the $p$-values for the remaining partitioning criteria are nonsignificant. This indicates that the items show different difficulties for respondents of different educational levels. An analysis of the $p$-values for the individual items reveals that in relation to the other items the probability word problems are much more difficult for respondents with an educational ISCED-level up to three.

In addition to the investigation of the assumption of person homogeneity we carried out an analysis with $T_{2}$ statistics to check for the possibility of the presence of item heterogeneity in the item set. The items were divided into two subgroups consisting of (1) rate-type algebra word problems and (2) probability-type algebra word problems.

Since the $T_{10}$ statistic for the partitioning criterion “educational level” reached the level of significance, we cal-

---

Table 1. Generative and quality control framework of the item generator AGEN (above) and an example item (below)

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic and grammatical complexity</td>
<td>Uncommon phrases and grammatically complex sentences are not used</td>
<td>Horn (1989)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Horn &amp; Noll (1997)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reusser (1989)</td>
</tr>
<tr>
<td>General word problem schema</td>
<td>Real world knowledge that is commonly but not necessarily assumed in algebra word problem solving is made explicit</td>
<td>Reusser &amp; Stebler (1997)</td>
</tr>
<tr>
<td>Presentational order</td>
<td>All word problems are presented in an “order naturalis”</td>
<td>Kintsch (1998)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Staub &amp; Reusser (1995)</td>
</tr>
<tr>
<td>Constructed response format</td>
<td>Reduce guessing probability</td>
<td>Enright et al. (2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sebrecht et al. (1996)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typicality of the cover story</td>
<td>The typicality of each cover story for all rate problem types is categorized as typical, neutral, or untypical.</td>
<td>Bernardo (1994)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Enright et al. (2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mayer (1981)</td>
</tr>
<tr>
<td>Number of partial equations</td>
<td>The total amount of partial equations</td>
<td>Enright et al. (2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hall et al. (1989)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lane (1991)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reed (1987)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Shalin &amp; Bee (1985)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sebrecht et al. (1996)</td>
</tr>
<tr>
<td>Number of relations between elements to be inferred</td>
<td>The relative amount of partial equations not explicitly mentioned in the problem text</td>
<td>Koedinger &amp; Nathan (2004)</td>
</tr>
<tr>
<td>Total number of unknown elements</td>
<td>Total amount of unknown elements in the solution equation</td>
<td>Enright et al. (2002)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hall et al. (1989)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Katz et al. (1996)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tabachneck et al. (1994)</td>
</tr>
</tbody>
</table>

Example item

<table>
<thead>
<tr>
<th>Radical</th>
<th>Categorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typicality of the Cover story</td>
<td>typical</td>
</tr>
<tr>
<td>Number of partial equations</td>
<td>4</td>
</tr>
<tr>
<td>Number of relations between elements to be inferred</td>
<td>1</td>
</tr>
<tr>
<td>Total number of unknown elements</td>
<td>2</td>
</tr>
</tbody>
</table>

A craftsman gets paid € 60,- during the first 3 hours of work. For each additional hour, he demands € 75,-. How many hours will he have to work, in order to earn € 255,-?
calculated the T2 statistics separately for respondents with an educational ISCED-level up to three and respondents with an educational ISCED-level of four or five. In the case of respondents with an educational level up to three, the T2 statistic results in a p-value of .0028 (Sim: 6000; Hits: 17), while the T2 statistic for respondents with an educational level of four to five results in a p-value of .048 (Sim: 6000; Hits: 288). Even though the deviation from the assumption of item homogeneity is more pronounced in the case of respondents with an educational level up to three, the T2 statistics are both statistically significant at \( \alpha = .05 \). Based on this result it seems reasonable to assume that algebra word problems classified as rate word problems and algebra word problems classified as probability word problems measure a *qualitatively different ability*, as these two algebra word-problem types belong to two different families of algebra word problems (Mayer, 1981).

Taken together, the results indicate that person homogeneity cannot be assumed for the combined rate and probability word-problem scale with regard to respondents’ educational level. Furthermore, the presence of item heterogeneity between rate and probability word problems points to qualitative differences in the processing of these two families of algebra word problems. Because of this result the implementation of a constraint regarding the algebra word-problem type seems to be justified. Based on these results, a decision was made to retain rate word-problem types, since this item type is one of the most commonly used word-problem types in current studies on cognitive processes in algebra word-problem solving.

### Study II

Study II was conducted to examine the dimensionality of a set of rate algebra word problems generated with AGen. A total of \( k = 15 \) rate word problems was generated using the set of four radicals outlined in Table 1. The number of partial equations varied from two to four (\( M = 2.8; \ SD = 0.68 \)), while the number of partial equations not explicitly mentioned in the problem text ranged from zero to two (\( M = 0.89; \ SD = 0.27 \)). Furthermore, the total number of unknown elements in the equation varied from one to five (\( M = 2.47; \ SD = 1.25 \)). Three different kinds of cover stories (typical, neutral, and atypical) were used to vary the typicality of the cover stories for the selected rate problems. The typicality of the cover story was determined individually for each template using the frequency norms reported by Mayer (1981). The main purpose of Study II was to generate a set of items covering a broad difficulty range and to investigate whether the built-in quality control framework of AGen suffices to generate algebra word-problem items at a high psychometric level.

The instructions given to the respondents was identical to the one described in the general procedure of this paper.

### Sample

The sample consisted of 147 (52.9%) male and 131 (47.1%) female respondents with ages ranging from 2 to 70 years (Mean=34.34; \ SD=12.55 \). A total of 18 (6.5\%) respondents completed less than 9 years of schooling (ISCED 1), 41 (14.7\%) respondents had completed 9 or 10 years of schooling without completing professional training (ISCED 2), 119 (42.8\%) respondents had completed professional training or technical school (ISCED 3), 59 (21.2\%) respondents had graduated from high-school (ISCED 4), and 41 (14.8\%) respondents had graduated from college or university (ISCED 5).

### Results

The reliability of the \( k = 15 \) automatically generated rate algebra items was a Cronbach’s \( \alpha \)-value of .87, indicating a sufficient degree of measurement precision from the viewpoint of classical test theory. We further investigated the psychometric property of this new item set using the Rasch model to investigate its one-dimensionality.

The assumption of person homogeneity was assessed through calculation and comparison of three MRM with one to three latent classes, respectively. Because of the low amount of some answer vectors, a parametric bootstrap (von Davier, 1996) was used to estimate the model-parameters. The results of the MRM analysis can be seen in Table 3.

### Table 3. Study II: Goodness-of-fit indices and information criteria for MRM with different number of latent classes for the scale “Algebraic Reasoning”

<table>
<thead>
<tr>
<th>Number of classes</th>
<th>( p (\text{Pearson } \chi^2) )</th>
<th>BIC</th>
<th>CAIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.062</td>
<td>1793.36</td>
<td>1809.36</td>
</tr>
<tr>
<td>2</td>
<td>.080</td>
<td>1815.57</td>
<td>1848.57</td>
</tr>
<tr>
<td>3</td>
<td>.020</td>
<td>1868.97</td>
<td>1918.97</td>
</tr>
</tbody>
</table>

Comparison with the saturated model.
As can be seen in Table 3, the one-class solution as well as the two-class solution does not deviate significantly from the saturated model. Using the information criteria CAIC (Consistent Akaike Information Criterion; Rost, 2004) and BIC (Bayesian Information Criterion; Rost, 2004), the one-class solution appears to be the more parsimonious model. Based on this result it can be assumed that there is no other partitioning of the sample that would result in a violation of the assumption of person homogeneity of the RM.

In addition, the assumption of item homogeneity was assessed by dividing the item set into two subgroups using the criteria (1) typicality of the cover story, (2) number of partial equations, and (3) total number of unknown elements in the equations. These three partitioning criteria were assumed to be the main radicals based on the theoretical model of algebra word-problem solving outlined above. With regard to the partitioning criteria “typicality of the cover story” the Martin-Löf statistic resulted in a \( \chi^2 \) value of 51.298 with \( df = 55 \) and \( p = .617 \). For the partitioning criteria “number of partial equations” the corresponding \( \chi^2 \) value amounts to 15.903 with \( df = 35 \) and \( p = .997 \), while the Martin-Löf statistic for the partitioning criteria “number of unknown elements in the equation” resulted in a \( \chi^2 \) value of 44.751 with \( df = 49 \) and \( p = .646 \). Thus, all three Martin-Löf statistics resulted in nonsignificant \( \chi^2 \) values; the corresponding effect size estimates suggested by Müller-Philipp and Tarnai (1989) all lie well below .30 (Min = .098, Max = .196). These results indicate that the assumption of item homogeneity can be retained. In sum, these results indicate that the 15 rate word problems used in this study indeed measure a one-dimensional latent trait.

### Study III

This study was carried out to investigate the nomothetic span (Embretson, 1983) of the automatically generated algebra word problems using confirmatory factor analysis. Based on the \( G_f-G_c \) theory (Horn, 1989; Horn & Noll, 1997) and the three-stratum theory (Carroll, 1993), we assumed that the observed correlations between these 10 subtests can be explained by the Cattell-Horn-Carroll model (CHC model), which assumes six factors and a higher-order \( g \)-factor. According to this theoretical model, the subtests Algebraic Reasoning and Computational Estimation should load on \( G_f \), while the subtests Figural-Inductive Reasoning and Verbal-Deductive Reasoning should load on \( G_v \). Verbal Comprehension and Verbal Production were assumed to constitute the factor \( G_c \), a single subtest, these two subtests were split into halves using an odd-even split procedure. Furthermore, the correlation between the latent factors should be explained by a higher-order \( g \)-factor. This model will be contrasted with a model that assumes that the correlations between the 10 subtests can be explained by a single latent factor. This model will be referred to as the \( g \)-factor model.

### Sample

The sample consisted of 101 (51.5%) male and 95 (48.5%) female respondents aged between 17 to 65 years (\( M = 37.74; SD = 11.94 \)). A total of 17 (8.5%) respondents had completed 9 years of school but no vocational training, while 75 (38.3%) respondents had also completed vocational school (ISCED 1 to 2). In addition, 84 (42.9%) respondents had a high school leaving certificate with university entrance permission (ISCED 3) and 21 (10.7%) respondents had graduated from university or college (ISCED 5 to 6).

### Measures

The test battery used in this study comprised two measures of quantitative reasoning (\( G_q \)), two measures of fluid intelligence (\( G_f \)), two measures of crystallized intelligence (\( G_c \)), two measures of short-term memory (\( G_{stm} \)), and one measure each for long-term memory (\( G_{ltm} \)) and visual processing (\( G_v \)). All subtests were taken from the INSBAT (Hornke et al., 2004), which has been shown to conform to the RM in previous studies (for a summary of the subtests used in the test battery as well as example items see Hornke et al., 2004; the test manual can be ordered free of charge at info@schufried.at). Thus, person parameters based on the RM will be used for each subtest in the analysis.

### Tested Models

Based on the \( G_f-G_c \) theory (Horn, 1989; Horn & Noll, 1997) and the three-stratum theory (Bickley et al., 1995; Carroll, 1993), we assumed that the observed correlations between these 10 subtests can be explained by the Cattell-Horn-Carroll model (CHC model), which assumes six factors and a higher-order \( g \)-factor. According to this theoretical model, the subtests Algebraic Reasoning and Computational Estimation should load on \( G_f \), while the subtests Figural-Inductive Reasoning and Verbal-Deductive Reasoning should load on \( G_v \). Verbal Comprehension and Verbal Production were assumed to constitute the factor \( G_c \), a single subtest, these two subtests were split into halves using an odd-even split procedure. Furthermore, the correlation between the latent factors should be explained by a higher-order \( g \)-factor. This model will be contrasted with a model that assumes that the correlations between the 10 subtests can be explained by a single latent factor. This model will be referred to as the \( g \)-factor model.
The CHC model provides a rather good fit to the data as indicated by all five global fit indices. None of the standardized residuals were statistically significant (< 1.96) at \( \alpha = .05 \). In contrast, none of the fit indices meet the standard criteria for an adequate model fit in the case of the g-factor model. It is, thus, hardly surprising that the CHC model fits the data significantly better than the g-factor model (\( \Delta \chi^2 [5] = 180.93; p < .10 \)).

The statistical significance of the factor loadings of the CHC model was also examined. The factor loadings of the individual subtests on the corresponding stratum-two factors were all moderate to high (\( \alpha > .60 \)) and reached statistical significance at \( \alpha = .01 \). The standardized factor loadings are depicted in Figure 1.

As can be seen in Figure 1, the subtest Algebraic Reasoning has the highest factor loading on \( G_q \) followed by Computational Estimation. The factor \( G_v \) is primarily defined by the subtest Figural-Inductive Reasoning followed by Visual Processing, while the \( G_c \) loads primarily on the subtest Verbal Comprehension followed by the subtest Verbal Production. The factor \( G_{stm} \) is marked by the subtest Verbal Short-Term Memory followed by Visual

---

**Table 4.** Study III: Descriptive statistics for all measures (above) and fit statistics for the hierarchical confirmatory factor analysis (below)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean</th>
<th>SD</th>
<th>Skew</th>
<th>Kurtosis</th>
<th>( r_e )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>−.14</td>
<td>0.72</td>
<td>.41</td>
<td>.39</td>
<td>.88(^a)</td>
</tr>
<tr>
<td>CE</td>
<td>−.52</td>
<td>1.66</td>
<td>−.60</td>
<td>1.10</td>
<td>.71(^a)</td>
</tr>
<tr>
<td>FID</td>
<td>−1.04</td>
<td>1.02</td>
<td>.18</td>
<td>.44</td>
<td>.70(^a)</td>
</tr>
<tr>
<td>VD</td>
<td>−.10</td>
<td>1.02</td>
<td>.82</td>
<td>1.63</td>
<td>.77(^a)</td>
</tr>
<tr>
<td>VC</td>
<td>.74</td>
<td>1.07</td>
<td>−.84</td>
<td>1.04</td>
<td>.75(^a)</td>
</tr>
<tr>
<td>VP</td>
<td>.02</td>
<td>1.31</td>
<td>−25</td>
<td>.20</td>
<td>.70(^a)</td>
</tr>
<tr>
<td>VI</td>
<td>.00</td>
<td>1.92</td>
<td>−.37</td>
<td>.91</td>
<td>.70(^a)</td>
</tr>
<tr>
<td>VE</td>
<td>−.39</td>
<td>1.05</td>
<td>.13</td>
<td>−.50</td>
<td>.70(^a)</td>
</tr>
<tr>
<td>RV</td>
<td>−.33</td>
<td>1.89</td>
<td>.28</td>
<td>−.18</td>
<td>.84(^a)</td>
</tr>
<tr>
<td>LM</td>
<td>−.78</td>
<td>1.02</td>
<td>.21</td>
<td>−.21</td>
<td>.72(^a)</td>
</tr>
</tbody>
</table>

\(^{a}\)Cronbach’s \( \alpha \), \(^{b}\)target reliability (CAT).
Short-Term Memory, while the factor loadings of the two parcels of the subtests Long-Term Memory and Spatial Comprehension were in close magnitude. With regard to the factor loadings of the g-factor on the stratum-two factors, the results indicate that the g-factor is nearly indistinguishable from $G_f$. The standardized factor loading of .99 is close to 1, and the error variance of $G_f$ did not reach the significance level ($\alpha > .05$). In order to test whether $G_f$ and the g-factor are indeed indistinguishable, the model was recalculated with the additional restriction that the standardized factor loading from the g-factor to $G_f$ equals 1. The restricted model yielded a $\chi^2 [48] = 63.58$ and $p = .065$. A $\Delta \chi^2$ test was used to investigate whether the restricted CHC model fits the data less well than the unrestricted model. The result ($\Delta \chi^2 [1] = 1.03; p = .311$) indicates that the restricted CHC model fits no worse than the unrestricted CHC model. It can, thus, be concluded that $G_f$ and the g-factor are indeed indistinguishable. This result is similar to the one obtained by Gustafsson (1984; Undheim & Gustafsson, 1987).

Taken together, the results of this study confirm the hypothesis that the subtest Algebraic Reasoning represents a marker of the second stratum factor $G_q$ and, thus, indicates that the nomothetic span of the new item material can be assumed.

Discussion

The present paper investigated whether algebra word problems can be automatically generated using an extended schema-based approach to automatic item generation, which includes a quality control framework to enhance the likelihood that the automatically generated algebra word problems are in accordance with the RM (Rasch, 1980) and, thus, have a high psychometric quality.

Based on a review of current approaches to automated item generation, the schema-based approach utilized in this paper was presented as an instance of an Approach 2 type (Dennis et al., 2002), which starts from a set of pre-existing templates for the construction of item isomorphs featuring the same item characteristics (radicals) as their “mother-items.” Following Arendasy (in press; Arendasy & Sommer, 2005), this schema-based approach was extended by incorporating a quality control framework in the item generator. The quality control framework essentially consists of a set of item features/rules that are assumed to (1) be associated with cognitive processes not related to the latent trait to be measured, or (2) possibly result in a differential item functioning. In the schema-based approach to automatic item generation these constraints were directly embedded into the templates.

The present series of studies was thus designed to investigate (1) the necessity of constraints regarding the inclusion or exclusion of different algebra word-problem families into the templates, (2) the efficiency of the constraints built into the quality control framework, and (3) the nomothetic span of the automatically generated algebra word problems. In Study I a set of eight rate word problems was automatically generated and presented together with eight probability word problems in order to investigate the necessity of a constraint regarding the restriction to a single algebra word-problem family (Mayer, 1981). The results of the nonparametric goodness-of-fit statistics for the RM indicated that person homogeneity cannot be assumed for the combined rate and probability word-problem scale with regard to the partitioning criterion “educational level.” A further analysis of the $p$-values for the individual items revealed that the violation of the assumption of person homogeneity regarding the partitioning criterion “educational level” can be attributed to differences in the difficulty of the probability word problems compared to the other algebra word problems. This means that probability word problems do not provide an “education-fair” measure of respondents’ algebraic reasoning ability. Furthermore, the presence of item heterogeneity between rate and probability word problems within each of the two educational level groups indicated qualitative differences in the processing of such items. This result underlines the importance of the implementation of a quality control framework into automated item generation as first suggested by Arendasy (2004, 2005, in press). The results of Study I can, thus, be seen as a validation of one of the constraints of the quality control framework built into AGen.

In Study II a set of $k = 15$ algebra word problems, which can be classified as rate word problems, was automatically generated using a set of radicals and a set of constraints derived from the current literature on algebra word-problem solving. Both the results of the MRM-based analyses as well as the Martin-Löf statistics indicated that the quality control framework seems to be rather effective in terms of automatically generating algebra word problems at a satisfactory psychometric level.

In order to investigate whether the implementation of the quality control framework affects the construct validity of the automatically generated items by restricting the possible range of algebra word-problem types to the rate word-problem family, a third study was conducted. The goal was to investigate the nomothetic span of the products of AGen according to the modified $G_f$-$G_c$ theory and the latest version of the three-stratum theory of intelligence. Using a hierarchical confirmatory factor analysis, the result of Study III supports the assumption that the automatically generated algebra word problems load on $G_q$ when administered together with another $G_q$-marker. The markers of the stratum-two factors $G_i$, $G_c$, $G_v$, $G_{ltm}$, and $G_{stm}$ provide evidence for the discriminant validity of the items. It was, thus, concluded that the implementation of a quality control framework does not necessarily reduce the construct validity of the automatically generated algebra word problems.

Taken together, the three pilot studies reported in this article highlight the importance of the implementation of a quality control framework into automated item generation.
as first suggested by Arendasy (2004, 2005, in press). However, the authors acknowledge that further studies need to be carried out in order to (1) examine further aspects of the construct representation of the automatically generated algebra word problems, and (2) provide further empirical support for the necessity of other possible constraints in AGen. From a practical point of view a schema-based approach to automated item generation based on a theoretically sound and empirically validated quality control framework would have several advantages: (1) The use of templates incorporating a unique set of radicals and constraints for the automatic generation of items helps to avoid misinterpretations in item writing in case a human item writer has to produce items above her or his ability level even when applying a set a generative rules. (2) The theory based definition of radicals and constraints enables a theoretically more sound interpretation of test scores, since the latent trait to be measured by the items is well defined by a generative framework. (3) The ability to generate new items on demand helps to increase test security, which is especially important for mass testing (e.g., college and university entrance exams).

References


Martin Arendasy
Faculty of Psychology
University of Vienna
Liebiggasse 5
A-1010 Vienna
Austria
Tel. +43 1 4277 47848
E-mail martin.arendasy@univie.ac.at